Bayesian profiling multiple imputation for missing hemoglobin values in electronic health records

Yajuan Si[1,*], **Mari Palta**[2], **Maureen Smith**[2]

[1]University of Michigan, Ann Arbor, Michigan, U.S.A.

[2]University of Wisconsin-Madison, Madison, Wisconsin, U.S.A.

Abstract

Electronic health records (EHRs) are increasingly used for clinical and comparative effectiveness research, but suffer from missing data. Motivated by health services research on diabetes care, we seek to increase the quality of EHRs by focusing on missing values of longitudinal glycosylated hemoglobin (A1c), a key risk factor for diabetes complications and adverse events. Under the framework of multiple imputation (MI), we propose an individualized Bayesian latent profiling approach to capture A1c measurement trajectories subject to missingness. The proposed method is applied to EHRs of adult patients with diabetes in a large academic Midwestern health system between 2003 and 2013 and had Medicare A and B coverage. We combine MI inferences to evaluate the association of A1c levels with the incidence of acute adverse health events and examine patient heterogeneity across identified patient profiles. We investigate different missingness mechanisms and perform imputation diagnostics. Our approach is computationally efficient and fits flexible models that provide useful clinical insights.

Keywords

Trajectory; Latent profile; Multiple imputation; Sensitivity analysis

1 Introduction

1.1 Glycemic testing and control

Diabetes is a condition requiring intensive management, and a major cause of morbidity and mortality. Approximately 25.2% of American seniors have diabetes and would benefit from individualized guidelines on glycemic testing and control (The Centers for Disease Control and Prevention 2017), but little clinical evidence exists to develop such plans. Current evidence-based guidelines encouraging tight glycemic control are most applicable to relatively healthy patients with diabetes. Glycemic control is usually monitored through glycosylated hemoglobin (A1c), which reflects blood glucose values over approximately the last 3 months. The American Diabetes Association (ADA) recommends A1c testing at least two times a year in patients who are meeting treatment goals and have stable glycemic control, and quarterly in patients whose therapy has changed or who are not

---
[*]yajuan@umich.edu .

meeting glycemic goals (ADA 2018a). Individuals without diabetes have A1c values below 5.7%, and maintaining A1c below 7% is recommended for individuals with diabetes to avoid chronic diabetes complications (ADA 2018a).

For patients aged 65 years or over and those with comorbid conditions, adhering tightly to guidelines established for those younger and healthier, might be difficult and diminish the quality of life. Tight control can lead to acute adverse outcomes, such as hypoglycemic coma, seizures, falls, fractures, motor vehicle accidents, cardiovascular events, stroke, and acute renal failure (Allen et al. 2001; Boyd et al. 2005; Davis and Alonso 2004; Kennedy et al. 2002; Pogach et al. 2007; Qaseern et al. 2007; Schech et al. 2007; Shorr et al. 1997). For example, the ACCORD and ADVANCE trials find that lowering A1c values increases the risk of cardiovascular death (Dluhy and McMahon 2008; Martin et al. 2006). The relationship between A1c and adverse outcomes in such patients may be U-shaped, and less stringent A1c goals may be appropriate (ADA 2018b).

To supplement findings from clinical trials, electronic health records (EHRs) are increasingly a data source for clinical and comparative effectiveness research on improving health care (Agency for Healthcare Research and Quality 2013; Cebul et al. 2011). The EHR provides a longitudinal record of patient medical information that is maintained by encounters in any care delivery setting, and includes key clinical data relevant to patient care, such as laboratory data (e.g., A1c values), demographics, progress notes, problems, medications, vital signs, past medical history, immunizations and radiology reports. Linked insurance data, for example from Medicare (U.S. Center for Medicare & Medicaid Services 2012), provide information on adverse clinical events through billed claims. However, the quality of EHRs is reduced by a large amount of intermittent missing data, because data are typically collected in an unscheduled fashion when the patient seeks care or the physician orders care.

In this article, we aim to use EHR to further understand the relationship between A1c levels and acute adverse outcomes. Information on adverse outcomes is rarely missing, as it is obtained from billing records. However, A1c within 3 months prior to an event may not be available due to the above recommendations. Multiple additional factors could influence the number of missing A1c values for individual patients. Physician non-adherence to guidelines for lab testing will cause missing values. The patient's health potentially affects the propensity to test, leading to informative missingness patterns that may cause systematic estimation bias. As argued by Haneuse and Daniels (2016), the complex interplay between health care providers and patients could result in various missing data mechanisms. Incorporation of all available information including patient health status is critical to appropriately handle missing A1c values from EHRs. Furthermore, capturing A1c trajectories and their relationship to patient characteristics is of clinical interest for health risk prediction and guideline establishment.

We propose a Bayesian profiling approach to incorporate patient characteristics and infer latent groups of A1c measurement trajectories. The measurement trajectories reflect both the A1c collection patterns and A1c values across time. The Bayesian profiling approach is combined with multiple imputation (MI, Rubin (1987)) to produce complete datasets for

general analysis purposes. As an illustration of the MI inference, we evaluate the association between A1c levels and the incidence of any acute health events, such as hospitalization, emergency room (ER) visits or death. The main novelty lies in generalizing MI with a flexible imputation engine, a covariate-dependent latent profile model, to depict nonlinear longitudinal trajectories and incorporate numerous covariates into the latent profiling.

1.2 Statistical literature on intermittently missing data

Missing A1c values in EHRs present statistical methodology challenges. Intermittent missing data in large-scale, unbalanced longitudinal studies, where subjects reappear after one or more missed visits, call for new imputation approaches especially when the missing values are potentially nonignorable (Little and Rubin 2002) and the observed cases are sparse. Simple methods (e.g., complete case analysis or last observation carried forward) will distort the relationship and are not recommended by the National Research Council (2010). Standard missing data methods in longitudinal studies focus on a common set of pre-specified and monotone missing times, where a measurement being missing implies that all follow-up measurements are also missing or dropped from the analysis (e.g., reviews in Ibrahim and Molenberghs (2009)). Likelihood-based, weighting or factorization approaches for missing data mainly apply to monotone missingness (Daniels and Hogan 2008). Utilization of the information collected after subjects reappear will potentially correct bias and increase estimation efficiency, and the challenge lies in how to appropriately model sparse data structures with non-monotone missing patterns. Weighting adjustment for longitudinal data with intermittent missingness is complex and computationally demanding, and there is no consistent recommendation on how the weights should be included in the longitudinal data modeling (Little and Rubin 2002). Sun and Tchetgen (2016) consider inverse probability weighting for non-monotone missing at random (MAR) data, whereas we argue that MI provides a superior solution to coherently utilize all available information and offer flexibility for model building with broad analytic goals.

We seek a flexible MI engine to predict intermittent missing values. MI separates imputation and analysis into two steps and propagates the uncertainty due to missing data. Various MI software packages have been developed assuming data are MAR and based on either joint multivariate normal distributions, e.g., PROC MI (SAS Institute Inc. 2017), Amelia (King et al. 2001) and norm (Schafer 1997), or a sequence of fully conditional distributions, e.g., IVEware (Raghunathan et al. 2001), mice (Van Buuren and Oudshoorn 1999), and mi (Gelman et al. 2015). Multilevel models are embedded with MI to handle correlated data, such as the packages REALCOM-IMPUTE (Carpenter and Kenward 2013) and pan (Schafer 2016). However, existing MI software cannot handle high-dimensional data that are subject to high proportions of intermittent missingness in large-scale longitudinal studies, nor nonignorable missing values (Si and Reiter 2013).

MI for missing not at random (MNAR) data requires a joint model for the incomplete variables (i.e., the A1c values) and the missingness indicators (i.e., whether A1c values are present on any given occasion). Examples of joint models include selection models, pattern-mixture models, and shared parameter models (Little 1995; Wu and Carroll 1988). Assumptions have to be introduced for parameter identification, which cannot be verified

with the available observations, and the computation is non-trivial (Ibrahim et al. 2005). The longitudinal A1c collection history in EHRs results in sparse observations across a large number of different missingness patterns and calls for flexible modeling strategies that account for the time dependency, borrow information across patterns and capture patient heterogeneity.

We jointly model the A1c missingness patterns and the lab values taking into account patient characteristics. To achieve parameter identification, dimension reduction and straight-forward interpretation, we introduce latent profiles and develop a Bayesian profiling multiple imputation (BPMI) approach. The novel latent profiling can handle ignorable and nonignorable missingness under the conditional independence assumption given the latent profiles, which yields a consistent estimation of parameters that are not associated with the latent profiles (McCulloch et al. 2016).

We assume that latent profiles are primarily determined by the trajectories of the longitudinal A1c measurements, including A1c values and potentially the measurement patterns. The latent profiling includes patient characteristics—such as socio-demographics, healthcare utilization measures, medications, medical complexity indicators, comorbidity and complications—as covariates that affect the profile assignment. This is an improvement over previous work that required joint modeling of the outcome variables, missingness indicators and covariates, leading to the need for Monte Carlo integration or approximate inferences (e.g., Zeldow et al. (2018)). The proposal shares a similar decomposition with Lin et al. (2004) but differs by the MI framework under the Bayesian paradigm. Besides avoiding the extra computation step of Monte Carlo integration, the posterior updating of BPMI has efficiency gains via Gibbs sampling and generates completed datasets for general analysis purposes.

As important contributions, the paper uses all available information with intermittent missingness to infer latent groups that are relevant to substantive interpretations, fits flexible mixture models that can capture irregular data distributions and develops Gibbs samplers to achieve computational efficiency. We examine the identified profiles and present the profile-specific characteristics to describe the subgroup heterogeneity.

The paper is organized as follows. Section 2 describes the EHR structure and content. The BPMI method is presented in detail in Section 3. We compare the new method with alternative imputation methods via simulation studies in Section 4 and apply it to the EHRs in Section 5, where our statistical and substantive findings are presented. Section 6 presents contributions and discusses future extensions.

2   EHR description

We extract the EHR data of adult patients with diabetes who belonged to a large academic Midwestern health system between 2003 and 2013. Patients are included if they have diabetes defined by a qualifying International Classification of Diseases-9 code for diabetes. We treat the first four quarters after the initial enrollment in the health care system as the baseline period (time 0). For inclusion, patients must have baseline A1c values and at least

one available follow-up A1c measurement. Patients are excluded from the analysis if they do not have continuous Medicare Part A & B fee-for-service, or if they are not seen at a clinic located in the system with access to laboratory data.

Because A1c reflects mean blood glucose over the preceding three months and quarterly testing is recommended for many patients, we construct a longitudinal dataset with one measurement per patient per three-month period (patient-quarter). When there are multiple A1c values available during the three months, we use the average of the measurements before the date of first acute health event (e.g., hospitalization, ER visit or death) date if there are any such events.

Missingness occurs when no measurements are available, or as occurs in a handful of cases (0.3%) after the first acute event in the quarter. We use the average of collected A1c values during the first four quarters as the baseline A1c value. Patient eligibility for inclusion is evaluated each quarter and patients are followed until loss to follow-up (e.g., no longer in the health system) or death, where for patients who die during the study period, we keep the patient in the dataset until after the quarter of death. Patients' various enrollment dates and lengths of follow-up result in an unbalanced data structure. We identified 7372 patients with 113761 quarters after baseline and only 57285 available observations. The proportions of intermittent missingness across patients are as high as 0.97, with a median of 0.50. The baseline A1c values of 7372 patients center around mean 6.93% and range from 3.55% to 17.1%, showing heterogeneity in glycemic control.

A "spaghetti" plot of A1c values is presented for 30 randomly selected patients in 2.1. Time trends and trajectories vary between patients, and repeated measurements of the same patients tend to be correlated. We consider mixed-effects models accounting for within-patient correlation with random intercepts and slopes with respect to time or functions of time. While the A1c measurement trajectories present patient heterogeneity, subgroups of patients could share similar profiles across time. We assume each subgroup to have different location and scale parameters for the time trends, which results in a mixture distribution overall.

Our goal is to estimate the A1c trajectories over time and identify latent A1c pattern groups, where the patient-level covariates affect the group allocation probabilities. The latent profiling allocation in the model takes into account variables that predict the A1c values or missingness patterns. EHRs collect a rich set of patient characteristics that could predict the A1c trajectories and likelihood of missingness patterns. The patient-quarter structure creates both time-invariant and time-varying variables.

Baseline characteristics include socio-demographics, healthcare utilization measures, medications, complexity, comorbid conditions (Elixhauser et al. 1998), and diabetes complications, all of which are carried forward (e.g., chronic conditions will be assumed as present from diagnosis onward). The comorbid patterns are characterized by Hierarchical Condition Categories (HCC) risk scores. The number of adverse events, such as hospitalization and ER visits, and the number of ER visits that did not lead to hospitalization during the first four quarters, serve as proxy measures of healthcare utilization. A large

number of variables and low prevalence of incidence cause problems in the covariate-dependent allocation, a well-known problem for latent class analysis (Vermunt 2010). We thus create a summary variable for each patient capturing the count of conditions with prevalence $< 2\%$. The covariates themselves could have high co-linear dependence and structural constraints. We use the observed A1c values ignoring the patient-quarter structure and fit an ordinary linear regression to select the variables that are highly correlated with A1c. We then fit a logistic regression model to predict the missingness at each quarter and select the variables that are predictive of the missingness pattern.

The union of the two sets of selected variables results in the covariate list in Table 2.1. The baseline characteristics are fully observed without missing values. The table shows that the cohort has a modest number of complex patients. For example, 14% of patients have Chronic Kidney Disease, 33% suffer from moderate or severe kidney damage, 19% have Congestive Heart Failure, 21% have Ischemic Heart Disease, and 85% have hypertension.

Time-varying variables include age and the count of physician visits that occurred within each quarter. On average patients have 2.39 physician visits every three months. We exclude time-varying Body Mass Index (BMI), blood pressure (BP), and elevated low-density lipoprotein (LDL), previously shown as associated with tight A1c control (Jackson et al. 2006; McFarlane et al. 2002; Niefeld et al. 2003) but subject to high missingness. LDL is often recommended for annual measurements. Measures such as BP and BMI are usually available when the patient sees a provider in a face-to-face visit, which occurs at A1c tests. The majority of BP records before 2007 in our dataset tend to be missing. Hence, the missingness percentages of these time-varying covariates are all above 50%, similar to those of A1c values.

Missing data in the time-varying covariates cause computational problems, and here we only consider missing A1c values and include other variables that are fully collected. The extension to handle missing values in the baseline and time-varying variables and perform a systematic variable selection by propagating all sources of uncertainty is discussed in Section 6.

3   Bayesian profiling multiple imputation

Denote the individual record for patient i by $\{X_{i0}, X_{ij}, Y_{ij}, R_{ij}, j \in [1, T_i]\}$ for the total number T_i of tracked quarters during the follow-up, where X_{i0} are the time-invariant covariates, X_{ij}'s are time-varying covariates, and Y_{ij} is the variable to be imputed, A1c values over time. For brevity, we will refer to Y_{ij} as the longitudinal outcome. Assume that only Y_{ij}'s are subject to missing values, and let R_{ij} be a time-varying binary indicator for its response, $R_{ij} = 1$ if Y_{ij} is observed, $R_{ij} = 0$ otherwise. Denote the time-varying variables by $Y_i = (Y_{i1}, \ldots, Y_{iT_i})^\top$, $X_i = (X_{i1}, \ldots, X_{iT_i})^\top$, and $R_i = (R_{i1}, \ldots, R_{iT_i})^\top$, for $i = 1, \ldots, n$ total number of patients.

Assume individuals fall into latent classes $C_i = 1, \ldots, L$, where L is a finite and positive known integer as the total number of classes. The selection of L and the case of L being an unknown random variable that induces additional uncertainty are discussed in Section 5 and Section 6. We assume that the latent class structure is primarily determined by the

trajectories of the longitudinal outcome and potentially its missingness pattern, and that the allocation probabilities are affected by the covariates.

The conditional distribution of the measurements given observed data can be expressed as

$$f(\mathbf{Y}_i \mid \mathbf{X}_{i0}, \mathbf{X}_i, \mathbf{R}_i = 1) = \sum_{l=1}^{L} \frac{f(\mathbf{R}_i = 1 \mid \mathbf{Y}_i, \mathbf{X}_{i0}, \mathbf{X}_i, C_i = l) f(\mathbf{Y}_i \mid \mathbf{X}_{i0}, \mathbf{X}_i, C_i = l) f(C_i = l \mid \mathbf{X}_{i0}, \mathbf{X}_i)}{f(\mathbf{R}_i = 1 \mid \mathbf{X}_{i0}, \mathbf{X}_i)},$$

where $f(\cdot)$ denotes the distribution.

If $\mathbf{R}_i$ only depends on fully observed $(\mathbf{X}_{i0}, \mathbf{X}_i)$, i.e., missingness is MAR, the observed data can provide valid inference since $f(\mathbf{Y}_i|\mathbf{X}_{i0}, \mathbf{X}_i, \mathbf{R}_i = 1) = f(\mathbf{Y}_i|\mathbf{X}_{i0}, \mathbf{X}_i)$. The profile structure is then independent of the missingness patterns.

$$f(\mathbf{Y}_i \mid \mathbf{X}_{i0}, \mathbf{X}_i, \mathbf{R}_i = 1) = \sum_{l=1}^{L} f(\mathbf{Y}_i \mid \mathbf{X}_{i0}, \mathbf{X}_i, C_i = l) f(C_i = l \mid \mathbf{X}_{i0}, \mathbf{X}_i).$$

We call this marginal profiling because the latent class structure influences only the marginal distribution of Y. The MAR assumption can be relaxed to

$$f(\mathbf{R}_i = 1 \mid \mathbf{Y}_i, \mathbf{X}_{i0}, \mathbf{X}_i, C_i = l) = f(\mathbf{R}_i = 1 \mid \mathbf{X}_{i0}, \mathbf{X}_i, C_i = l), \tag{3.1}$$

that is, conditional MAR, where the missingness is independent of the outcome given the latent classes and covariates. However, the unconditional missingness is nonignorable as the latent structure affects both the longitudinal outcome and missingness patterns. Joint modeling of $(\mathbf{Y}_i, \mathbf{R}_i)$ is then necessary and will be identified due to the conditional independence assumption given the latent classes. We call this approach joint profiling.

Joint profiling is identical to marginal profiling when the parameters in Model (3.1) do not change across profiles, so $\mathbf{R}_i$ is conditionally independent of C_i given $(\mathbf{X}_{i0}, \mathbf{X}_i)$. Hence, the joint profiling approach can model both ignorable and nonignorable missing data. We will consider both the marginal and joint profiling for imputation and make inferences on the effect of A1c levels on acute health outcomes.

The latent classes are at the individual level and thus time invariant. Further, we assume the latent class allocation depends only on time-invariant variables $f(C_i = l|\mathbf{X}_{i0}, \mathbf{X}_i) = f(C_i = l|\mathbf{X}_{i0})$. For computational and interpretational convenience, we assume that the outcome trajectory is independent of time-invariant covariates given the latent class $f(\mathbf{Y}_i|\mathbf{X}_i, \mathbf{X}_{i0}, C_i = l) = f(\mathbf{Y}_i|\mathbf{X}_i, C_i = l)$. That implies that the patient characteristics at baseline are represented by the latent profiling.

Here, the conditional independence assumption between Y_{ij} and R_{ij} given C_i is used for identification, and the conditional independence assumptions involving X_{i0} and X_{ij} are introduced for interpretation and can be tested in the model fitting.

Are our assumptions regarding the missing data too strong? Haneuse and Daniels (2016) list various sub-mechanisms—potentially MAR or MNAR—that are relevant to missing EHRs, such as enrollment status, encounters with the health system, measurements and structural changes. McCulloch et al. (2016) evaluate biased and unbiased estimation in longitudinal studies with informative missingness and recommend mixed-effects models. In addition to random effects, we introduce latent profiles with a large number of covariates, and expect the flexible latent profiling with rich characteristics to capture the underlying heterogeneity and make the assumptions at least plausible. We rigorously assess the imputation performances in Section 5.

3.1 Marginal profiling

Under MAR we construct latent profile models based on the collected outcomes, i.e., marginal profiling, which is estimated from the pattern of trajectories of the longitudinal outcomes.

$$f(\boldsymbol{Y}_i^{obs} \mid \boldsymbol{X}_{i0}, \boldsymbol{X}_i) = \sum_{l=1}^{L} f(\boldsymbol{Y}_i^{obs} \mid \boldsymbol{X}_i, C_i = l) f(C_i = l \mid \boldsymbol{X}_{i0}). \tag{3.2}$$

Here $\boldsymbol{Y}_i^{obs}$ denotes the observed A1c measurement for individual i. The factorization involves an outcome model and a latent profile model. The model specification in (3.2) assumes: 1) the missingness pattern is independent of the A1c values and the latent profiles given the covariates $R_{ij} \perp\!\!\!\perp (Y_{ij}, C_i) | (\boldsymbol{X}_{i0}, \boldsymbol{X}_{ij})$; 2) the A1c values are independent of the missingness patterns and the baseline characteristics given the latent profiles and time-varying covariates $Y_{ij} \perp\!\!\!\perp (R_{ij}, \boldsymbol{X}_{i0}) | (C_i, \boldsymbol{X}_{ij})$; and 3) the latent profiles are independent of the time-varying covariates given the baseline characteristics $C_i \perp\!\!\!\perp \boldsymbol{X}_{ij} | \boldsymbol{X}_{i0}$. Since the role of $\boldsymbol{X}_{i0}$ can be checked, Model (3.2) can be generalized by assuming the outcome depends on the baseline characteristics and that these are captured by the latent profiling in the current specification. Our sensitivity analysis in the EHR application does not find evidence against the specification with $\boldsymbol{X}_{i0}$, as the covariates affect the allocation assignment probability of different profiles.

In the latent profile model, denote by $f(C_i = l | \boldsymbol{X}_{i0}) \doteq \pi_l(\boldsymbol{X}_{i0})$ the allocation probability of pattern l conditional on $\boldsymbol{X}_{i0}$, for $l = 1, \dots, L$. We consider a multinomial logistic regression model with coefficient vector $\boldsymbol{\eta}_l$, for $l = 1, \dots, L$. Set $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_L)$ and $\boldsymbol{\eta}_1 = \vec{0}$ for identification, we have

$$\pi_l(\boldsymbol{X}_{i0}) = \frac{\exp(\boldsymbol{X}_{i0}^T \boldsymbol{\eta}_l)}{\sum_{k=1}^{L} \exp(\boldsymbol{X}_{i0}^T \boldsymbol{\eta}_k)}. \tag{3.3}$$

For posterior computation we develop a Gibbs sampler by introducing Pólya-Gamma (PG) distributed variables w_{il} (Polson et al. 2013),

$$w_{il} \mid (\boldsymbol{X}_{i0}, \boldsymbol{\eta}) \sim \text{PG}(1, r_{il}),$$

where $r_{il} = X_{i0}^T \eta_l - \log[\sum_{k \neq l} \exp(X_{i0}^T \eta_k)]$. Conditional on the PG variables, the posterior distribution of η_l will be conjugate with normal prior distributions. The resulting Gibbs sampler improves the posterior fitting and outperforms rejection sampling methods with quick convergence.

Let $c_{il} = I(C_i = l)$ be the latent class indicator and introduce the normal prior distribution $N(\mathbf{b}_l, \mathbf{B}_l)$ on the coefficients η_l, the conditional posterior distribution of η_l given w is multivariate normal, $\pi(\eta_l \mid -) \sim N(\mathbf{m}_l^*, \mathbf{S}_l^*)$. Here $\mathbf{S}_l^* = (\mathbf{V}^T \Omega_l \mathbf{V} + \mathbf{B}_l^{-1})^{-1}$, $\mathbf{V}$ is the design matrix with each row X_{i0}^T, $\Omega_l = diag(w_{1l}, \ldots, w_{nl})$, and $\mathbf{m}_l^* = \mathbf{S}_l^*(\mathbf{B}_l^{-1}\mathbf{b}_l + \mathbf{S}^T\mathbf{m}_l)$, where $\mathbf{m}_l$ is a vector in R^n, and the ith component is $m_l^{(i)} = c_{il} - 1/2 + w_{il}\{\log[\sum_{k \neq l} \exp(X_{i0}^T \eta_k)]\}$.

For the collected longitudinal and continuous A1c measurements, we assume a linear mixed-effects model with profile-specific coefficients and variances (β_C^*, σ_C^2). Let $\boldsymbol{D}_i$ denote a $T_i \times p$ covariate matrix, with associated p-vector of coefficients β. The jth row of $\boldsymbol{D}_i$, denoted by $\boldsymbol{D}_{ij}$, is then a p-vector of covariate values measured at time j. Covariates for profile-specific effects $\beta_{C_i}^*$ and for individual-specific random effects $\mathbf{b}_i$ are denoted by the $T_i \times q$ matrix $\boldsymbol{D}_i^*$ and the $T_i \times r$ matrix $\boldsymbol{D}_i^{**}$, respectively, both with structures similar to $\boldsymbol{D}_i$. There may be overlap of the covariates in $\boldsymbol{D}_i$, $\boldsymbol{D}_i^*$ and $\boldsymbol{D}_i^{**}$, including main effects and high-order interactions of X_i. Deterministic functions of time, for example, basis spline functions of time in our EHR application, can be included in the covariates.

$$Y_i = \boldsymbol{D}_i\beta + \boldsymbol{D}_i^*\beta_{C_i}^* + \boldsymbol{D}_i^{**}\mathbf{b}_i + \epsilon_i$$
$$\mathbf{b}_i \sim N(\overrightarrow{0}, \Sigma_{r \times r}), \text{ and } \epsilon_i \sim N(\overrightarrow{0}, \sigma_{C_i}^2 I_{T_i \times T_i}),$$

(3.4)

which is a linear mixed-effects model with a mixture of location and scale parameters varying across profiles. The profile-specific coefficients $\beta_{C_i}^*$ can capture the differential trajectories of A1c measurements with time-varying covariates in $\boldsymbol{D}_i^*$. Model (3.4) is a location and scale mixture model that is flexible enough to capture non-Gaussian distributions as shown in Figure 2.1. Here $\Sigma_{r \times r}$ is the covariance-variance matrix of the individual-specific random effects $\mathbf{b}_i$, which could be simplified as a diagonal matrix if the components of $\mathbf{b}_i$ are treated as independent or a scalar if $\mathbf{b}_i$ only represents random intercepts.

We assign weakly informative and conjugate prior distributions to the parameters (Gelman et al. 2008). The prior specification and full conditional posterior distributions as efficient Gibbs sampler are presented in Appendix B.1. After convergence, based on the posterior samples of the parameters, we can impute the missing A1c values and disseminate completed datasets.

3.2 Joint profiling

With MNAR, we jointly model the A1c values and missingness patterns given the covariates to obtain the joint profiling structure

$$f(\boldsymbol{Y}_i, \boldsymbol{R}_i \mid \boldsymbol{X}_{i0}, \boldsymbol{X}_i) = \sum_{l=1}^{L} f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, C_i = l) f(\boldsymbol{R}_i \mid \boldsymbol{X}_i, C_i = l) f(C_i = l \mid \boldsymbol{X}_{i0}), \qquad (3.5)$$

where $\boldsymbol{Y}_i$ includes both observed and missing measurements. The joint model has three components: the latent profile model, the longitudinal outcome model and the longitudinal response propensity model.

This model captures the MNAR dependence between the outcome and missingness patterns that jointly determine the latent profiles. The response indicator is conditionally independent of the outcome given the latent profiles and covariates, i.e., conditional MAR: $R_{ij} \perp\!\!\!\perp (Y_{ij}, X_{i0}) \mid (C_i, X_{ij})$. Conditional MAR makes models identifiable, yet flexible for dimension reduction and interpretation.

We consider a generalized linear mixed-effects model with a logit link for the longitudinal response indicator R_{ij}, conditional on latent classes and random effects $\mathbf{e}_i \sim N(0, \mathbf{E})$ with $\mathbf{E}$ as a covariance-variance matrix.

$$\mathrm{logit}\mathrm{Pr}(R_{ij} = 1 \mid \boldsymbol{X}_{ij}, C_i, \mathbf{e}_i)) = \boldsymbol{D}_{ij}\boldsymbol{\alpha} + \boldsymbol{D}_{ij}^{*}\boldsymbol{\alpha}_{C_i}^{*} + \boldsymbol{D}_{ij}^{*\ *}\mathbf{e}_i, \qquad (3.6)$$

where the covariates $(\boldsymbol{D}_{ij}, \boldsymbol{D}_{ij}^{*}, \boldsymbol{D}_{ij}^{*\ *})$ could overlap or be different from those selected in Model (3.4) as subsets of the main effects and high-order interactions in $\boldsymbol{X}_{ij}$.

3.3 Summary

The full model specification under marginal profiling is summarized below:

$$\begin{aligned}
&\boldsymbol{Y}_i \sim N(\boldsymbol{D}_i\boldsymbol{\beta} + \boldsymbol{D}_i^{*}\boldsymbol{\beta}_{C_i}^{*} + \boldsymbol{D}_i^{*\ *}\mathbf{b}_i, \sigma_{C_i}^2 I_{T_i \times T_i}) \\
&C_i \sim \mathrm{Multinomial}(\pi_1(\boldsymbol{X}_{i0}), \ldots, \pi_L(\boldsymbol{X}_{i0})) \\
&\quad \pi_l(\boldsymbol{X}_{i0}) = \frac{\exp(\boldsymbol{X}_{i0}^T\boldsymbol{\eta}_l)}{\sum_{k=1}^{L} \exp(\boldsymbol{X}_{i0}^T\boldsymbol{\eta}_k)} \\
&\mathbf{b}_i \sim N(\overrightarrow{0}, \boldsymbol{\Sigma}_{r \times r}), \ \beta \sim N(0, \boldsymbol{\Sigma}_\beta), \ \boldsymbol{\Sigma}_{r \times r} \sim \mathrm{Inverse\text{-}Wishart}(\nu_b, \boldsymbol{\Sigma}_b) \\
&\sigma_l^2 \sim \mathrm{Inverse\text{-}Gamma}(a, b), \ \eta_l \sim N(b_l, B_l), \ \text{for } l = 1, \ldots, L.
\end{aligned} \qquad (3.7)$$

The full model specification under joint profiling is summarized below:

$$Y_i \sim N(D_i\beta + D_i^*\beta_{C_i}^* + D_i^{*\,*}\mathbf{b}_i, \sigma_{C_i}^2 I_{T_i \times T_i})$$

$$R_{ij} \sim \text{Bernoulli}(\Pr(R_{ij} = 1))$$

$$\text{Inverse-logit}(\Pr(R_{ij} = 1)) = D_{ij}\alpha + D_{ij}^*\alpha_{C_i}^* + D_{ij}^{*\,*}\mathbf{e}_i$$

$$C_i \sim \text{Multinomial}(\pi_1(X_{i0}), \dots, \pi_L(X_{i0}))$$

$$\pi_l(X_{i0}) = \frac{\exp(X_{i0}^T\eta_l)}{\sum_{k=1}^{L}\exp(X_{i0}^T\eta_k)} \tag{3.8}$$

$$\mathbf{b}_i \sim N(\overrightarrow{0}, \Sigma_{r \times r}),\ \beta \sim N(0, \Sigma_\beta),\ \Sigma_{r \times r} \sim \text{Inverse-Wishart}(\nu_b, \Sigma_b)$$

$$\alpha_l \sim N(0, \Sigma_\alpha),\ \eta_l \sim N(b_l, B_l),\ \sigma_l^2 \sim \text{Inverse-Gamma}(a, b),$$

$$\nu \sim N(0, \Sigma_\nu),\ \gamma_l \sim N(0, \Sigma_\gamma),\ \mathbf{e}_i \sim N(0, E),\ E \sim \text{Inverse-Wishart}(\nu_e, \Sigma_e).$$

We set the hyper-parameters equal to 1 and specify an independent covariance matrix structure. Our inference is not sensitive to the specification of hyper-parameter values (e.g., 0.1, 0.01) under noninformative or weakly informative prior settings, as expected given the large sample sizes in the application. The PG variables w_{il}'s are introduced for posterior updating with η. We develop the Gibbs sampler with two sets of introduced Pólya-Gamma distributed variables for posterior computation with models (3.3), (3.4) and (3.6) under weakly informative and conjugate prior distributions. The imputation of missing A1c values is nested in the iterative process, unlike marginal profiling, where parameter estimation and data augmentation are implemented simultaneously. Computational details are presented in Appendix B.2.

For ease of interpretation of the potential profiling structure, we fix the number of latent patterns L to be a small integer, chosen with the aid of diagnostic tools for model selection described in Section 5.1. The model can be extended to allow nonparametric Bayesian modeling, such as the dependent probit stick-breaking process (Rodríguez and Dunson 2011). However, discrete covariates with low prevalence tend to cause separation problems with a large number of latent profiles, as in our application study.

The BPMI in (3.7) and (3.8) uses mixture models to fiexibly capture irregular distributions, incorporate time trends, and latent profiles allow us to jointly model the outcomes and nonignorable missingness patterns.

4   Simulation

We evaluate the imputation quality and MI inferential validity in repeated sampling. We simulate unbalanced longitudinal data, and the outcome is subject to intermittent missing values. For each sample, we generate $n = 500$ subjects with varying lengths of follow-up periods T_i's randomly selected from $\{1, \dots, 10\}$ and time-invariant covariates D_i with four binary indicators simulated from Bernoulli's distributions with probabilities 0.2, 0.3, 0.4, and 0.5, respectively. The time-varying covariate is the time after the first collection, where we use t_{ij} denoting the time points of collection divided by 4 to mimic the quarter structure in the EHR data. We consider different outcome generation models and different missingness mechanisms, as we now describe.

In Case 1, we groups the 500 subjects into $L = 3$ classes with allocation probabilities $\exp(D_i\eta_l)/(1 + \exp(D_i\eta_l))$, where $\eta_1 = \overrightarrow{0}$, $\eta_l = 0.5 * (1{:}5) - 0.5 * l$, for $l = 2, 3$. The outcome Y is simulated from a 3-component finite mixture of normal distributions with mean $\mu_{ij} = b_i + D_i\beta_0 + t_{ij}\beta_{1l} + t_{ij}^2\beta_{2l}$ and standard deviation (sd) σ_l, for $l = 1, 2, 3$, respectively. Here we set $\beta_0 = (2, 2, 2, 2, 2)$, $\beta_{11} = 0$, $\beta_{12} = -2$, $\beta_{13} = 2$, $\beta_{21} = 0$, $\beta_{22} = 1$, $\beta_{13} = 1$, $\sigma_1 = 0.1$, $\sigma_2 = 1$, and $\sigma_3 = 3$. We assume the intermittent missingness is MAR with the response probability $\Pr(R_{ij} = 1) = \exp(D_i a_0) + t_{ij})/(1 + \exp(D_i a_0 + t_{ij}))$ for $j = 2, \ldots, T_i, i = 1, \ldots n$, where $a_0 = (-2, -1, 0, 1, 2)$. The missingness mechanism depends on the main effects of the covariates and time. The missingness proportion is around 38%. The data generation and missingness mechanism in Case 1 are consistent with the outcome models and identification assumptions under marginal profiling, denoted as Mixture–Main.

In Case 2, the outcome is simulated from a normal distribution with mean $\mu_{ij} = b_i + D_i t_{ij}\beta_0 + t_{ij}\beta_1$ and sd=1, where the interaction term between the time-invariant covariates and time are included with $\beta_0 = (2, 2, 2, 2, 2)$ and $\beta_1 = 1$. The model assumes that the linear trends vary across individuals. The missingness mechanism is the same as that in Case 1. We denote the data generation and missingness mechanism in Case 2 as Interaction–Main, which differs from the outcome model assumptions under BPMI.

In Case 3, the outcome model has the same specification as that in Case 2, with interaction terms. The missingness also depends on the interaction terms between the time-invariant covariates and time: $\Pr(R_{ij} = 1) = \exp(e_i + D_i a_0 + t_{ij} a_1 + D_i t_{ij} a_2) / (1 + \exp(e_i + D_i a_0 + t_{ij} a_1 + D_i t_{ij} a_2))$ for $j = 2, \ldots, T_i, i = 1, \ldots n$, where $a_0 = (1, 1, 1, 1, 1)$, $a_1 = -0.5$, $a_2 = (-2, -1, 0, 1, 2)$, and $e_{ij} \sim N(0, 1)$, with resulted missingness around 20%. We denote the data generation and missingness mechanism in Case 4 as Interaction–Interaction, which differs from assumptions for both the outcome and missingness mechanism under BPMI.

In Case 4, the outcome model is the same as that in Case 1.
We assume that the missingness depends on the group structure, and it is conditionally MAR but marginally MNAR. The response probability is $\Pr(R_{ij} = 1) = \exp(e_i + D_i\alpha_0 + t_{ij}\alpha_{1l} + t_{ij}^2\alpha_{2l}) / (1 + \exp(e_i + D_i\alpha_0 + t_{ij}\alpha_{1l} + t_{ij}^2\alpha_{2l}))$ for $j = 2, \ldots, T_i, i = 1, \ldots n$, where $\alpha_0 = (1, 1, 1, 1, 1)$, $\alpha_{11} = 0$, $\alpha_{12} = -2$, $\alpha_{13} = 2$, $\alpha_{21} = 0$, $\alpha_{22} = -4$, $\alpha_{23} = 4$, and $e_{ij} \sim N(0, 1)$, resulting in around 35% of missing values. The scenario in Case 3 is similar to that under joint profiling, denoted as Mixture–Mixture.

In Case 5, we violate the conditional MAR assumption by specifying the response propensity still depends on the outcome given the group structure. The remaining specifications are the same as those in Case 4. This case is denoted as Mixture–Outcome.

We implement the marginal profiling approach (BPMI-MAR), the joint profiling approach (BPMI-MNAR), with the popular imputation methods MICE and multilevel imputation MICE-2l, to generate 10 imputed datasets each. Both MICE and MICE-2l assume MAR. MICE treats measures at the 10 time points as 10 different variables, performs fully conditional chained imputation with linear regression models and fills in missing values with draws from observed cases based on predictive mean matching by default. This accounts

flexibly for the correlation between repeated measures but ignores any pattern across time. MICE generates balanced data, where measures at all time occasions are filled in even though some are not eligible in the intrinsic data structure. MICE-2l fits a multilevel model for the outcome with a random intercept and a random slope for time. MICE-2l fits a linear time trend, and assumes normality of all random components. Hence, BPMI differs from the other two approaches in being able to handle MNAR, having more flexible distributional assumptions, modeling the time trend in a flexible manner and in handling covariates in an implicit manner with potential interaction terms rather than by an explicit inclusion of main effects.

We run BPMI-MAR in the exact specification in (3.7) for 3000 iterations and BPMI-MNAR in (3.8) for 2000 iterations with $L = 3$. The time-varying covariates in BPMI include the basis spline functions of time used in Section 5 and given in Appendix A. Our quantities of interest are the average values of the outcome in the follow-up period, across $j = 2, \ldots, 10$. We perform repeated sampling for 100 times and stack the generated samples treated as the population. We combine the MI inferences and calculate average values of the bias, root mean squared error (RMSE) and nominal coverage rate of 95% confidence intervals for each of these four approaches. We divide the bias and RMSE by the true value and present the relative bias (Rel–Bias) and relative RMSE (Rel–RMSE), and coverage rates in Figure 4.1.

The marginal profiling approach is the most robust and efficient method across the five cases. In Case 1, the sequential imputation MICE generates unbiased estimates but with larger RMSEs and smaller coverage rates than the BPMI-MAR. The performance of joint profiling BPMI-MNAR is competitive, with relatively larger bias and RMSE but negligible differences. The coverage rates of multilevel imputation models MICE-2l are generally lower than 0.95 even with large RMSEs.

Case 2 shows that BPMI can recover the interaction effects through latent profiling when the covariates enter through the allocation probabilities but are excluded in the outcome modeling and imputation. This is also confirmed in Case 3, when interaction terms are present in the response propensity model. In Case 4, with marginal MNAR and conditional MAR, MICE and MICE-2l fail to yield valid inferences as both methods assume MAR. The BPMI approaches outperform with small bias, RMSEs and reasonable coverage rates. The outputs of joint profiling BPMI-MNAR did not show improvement over those under marginal profiling. This could be due to the Monte Carlo error with not enough number of repeated samples or the computation problems of BPMI-MNAR that requires more iterations to converge. Both BPMI and MICE generate large biases and RMSEs in Case 5, when the missing data mechanism assumption is violated.

5 Imputation and Inference with EHRs

In a two-step MI process, we first apply BPMI to impute missing A1c values in the EHRs and secondly make inference on how A1c relates to acute diabetes outcomes. The observed A1c values range from 3.3% to 17.3% with mean 7.1% and are right-skewed with multiple modes. We use the location and scale mixture distribution as a flexible strategy to capture the irregular distribution. In the profile allocation model (3.3), the time-invariant

covariates X_{i0} are the baseline characteristics listed in Table 2.1. In the A1c outcome and response propensity models (3.4) and (3.6), the covariates in D_{ij} with fixed effects include the constant ones and the time-varying variables, age and the count of physician visits. The covariates in D_{ij}^* represent the functions of time in years with slopes changing across classes. We use the basis spline functions of time with (1%, 15%, 20%, 50%, 75%, 90%) quantiles as six knots and cubic terms as a piecewise polynomial regression, and the nine spline functions used as covariates in D_{ij}^* with profile-specific coefficients as shown in Appendix A. The spline functions are chosen based on model fitting criteria and can handle complex shapes to create smooth curves. Details on diagnostics of model specifications are provided in Section 5.1. Random intercepts are introduced by D_{ij}^{**}.

To set up the initial values of the Markov chain Monte Carlo (MCMC) updating, we fit an ordinary linear regression and use the coefficient estimates as initial values for β and β_l^*'s whose starting points are the same across classes. The initial values of η's are $\vec{0}$, representing an initial equal probability of class allocation. The scale parameters start at 0.1, and the random effects are initially drawn from normal distributions with mean 0 and the initial scales.

We implement posterior computation for both marginal and joint profiling. The MCMC algorithms efficiently achieve convergence with Gibbs samplers as diagnosed by trace-plots of the posterior samples, where 15000 iterations under marginal profiling and 5000 iterations under joint profiling take ~5 hours to finish each with R codes run on a standard laptop (3.5 GHz Dual-Core Intel Core i7, 16 GB Memory), and the computational speed can be substantively improved with R–C++ interface and parallel computing resources. We run the MCMC chains long enough to obtain 100 multiply imputed datasets and 500 replicated datasets to check if imputation is able to predict observed outcomes. For classification, we follow Goodman (2007) with a hard partitioning and keep one random assignment based on a random draw from the component-specific probabilities across iterations. Our analysis shows that the results are not sensitive to the classification rules.

5.1 Model diagnostics and comparison

We use the Bayesian information criterion (BIC) and the log-pseudo marginal likelihood (LPML) to select the number of classes, L. For BIC, we use the posterior mean estimates of related parameters and obtain the likelihood conditional on the class allocation. LPML is the sum of logarithms of conditional predictive ordinates and estimated using posterior samples of parameters, $\theta(t)$, $t = 1, \ldots, T$, and LPML $= [1 / T \sum_{t=1}^{T} 1 / f(Y_i \mid \theta^{(t)})]^{-1}$ (Gelfand and Dey 1994). This is an approximation of the leave-one-out cross-validation using importance sampling. We use $f(Y_i \mid \theta^{(t)})$ for the marginal profiling and $f(Y_i, R_i \mid \theta^{(t)})$ for the joint profiling.

We restrict L to be small to avoid computational problems due to the separation of the low-frequency predictors, such as dementia and renal failure. Table 5.1 presents the BIC and LPML values for models with different L values. The model with the smallest BIC and largest LPML will be favored. We see that the model with $L = 3$ classes presents a reasonable choice both for marginal and joint profiling.

To assess whether imputations are plausible, we perform a posterior predictive check and generate replicated datasets that are predicted values of observations based on the posterior estimates of model parameters after convergence (Meng 1994). Let $\{R^{(1)}, \dots, R^{(T^0)}\}$ be the collection of the T^0 replicated data sets. We then compare statistics of interest in each replicated dataset to those in the observed dataset D. Suppose that S is some statistic of interest, let $S_{R^{(t)}}$ and S_D be the values of S computed from $R^{(t)}$ and D, respectively. The quantities S include the mean, 2.5% percentile and 97.5% percentile of A1c levels for every patient and for every quarter. For the patient-level summaries, we look at the two-sided posterior predictive probabilities as diagnostic tools defined as

$$\text{ppp} = \frac{2}{T^0} * \min\left(\sum_{t=1}^{T^0} I(S_{R^{(t)}} - S_D > 0), \ \sum_{t=1}^{T^0} I(S_{R^{(t)}} - S_D < 0) \right).$$

When the value ppp is small, for example, less than 5%, this suggests that the replicated data sets systematically differ from the observed data set, with respect to that statistic. With larger ppp values, the evidence does not contradict that the imputation model preserves the observed characteristics of that statistic (He et al. 2010; Si et al. 2016). For the 7372 patient-level A1c summaries, the marginal profiling model yields 52 ppp values that are below 1% for the average, and the numbers of below 1% ppp values for the 2.5% percentile and the 97.5% percentile are 238 and 375. The number of ppp values that are below 1% for these three statistics under joint profiling is 66, 464 and 678, respectively. Based on the low proportions of small ppp values, the posterior predictive check does not indicate lack of model fit in recovering the trajectories. However, the performance of joint profiling is inferior to marginal profiling in preserving the observed mean values across patients.

For the quarter-level A1c summaries, we compare the predictions with the observed values under the two profiling approaches. The 95% predictive credible intervals greatly overlap with the empirical 95% confidence intervals (CIs). Figure 5.1 shows that the posterior predictive mean estimates are generally close to the observed mean values, except for two observations, and both approaches perform competitively and similarly. We check the parameter estimates in Model 3.1, and find that the coefficients do not change across profiles. This explains the similar performance of joint profiling and marginal profiling.

We also compare BPMI with MICE and MICE-2l, by examining the summaries of the imputed A1c values. Figure 5.2 depicts the average values and 95% confidence intervals of A1c measurements in the follow-up period from the 5th to the 44th quarter. We do not have a gold standard for the missing values and aim to check which method can recover the data structure. With the MAR assumption for MICE and BPMI under marginal profiling, the time dependence structure should be the same between the imputed data and observed data. Comparing the trajectories of the imputed data to those of the observed data in Figure 5.1, marginal profiling under BPMI yields imputation that presents the patterns most similar to the observed data, showing the trend of BPMI first flat and then increasing. The failure of MICE to impose a time structure results in wiggly averages and the last measurement having the lowest value. In addition, the need for MICE to impute occasions post drop-out or death may have led to inconsistent means in the later quarters. The multilevel imputation

MICE-2l smooths the variation across time and presents a linear trend, which is different from the observed trends in Figure 5.1. Both MICE approaches tend to generate larger imputed values than those under BPMI, potentially due to the ability of BPMI to better replicate a multimodal skewed distribution of A1c values.

In fact, BPMI imputes a substantially higher proportion of A1c values in the 6-7% range and fewer in the 7-8% range or below 6% than do MICE or MICE-2l. Based on a randomly selected single imputed dataset, 18.3% of the imputed values from BLPM are below 6%, 44% values are between 6% and 7%, and 23% values are in the 7-8% range. However, the imputed values from MICE have 20% below 6%, 37% between 6% and 7%, and 27% in the 7-8% range, and the percentages from MICE-2l are 21%, 36% and 28%, respectively. Having more imputed data in the below 7% range resonates clinically with the fact that patients in that range have met ADA goals and may be seen as needing less frequent A1c testing.

5.2   Inference and profiling

After MI, we analyze the completed datasets to evaluate the association of A1c with acute health event incidence and depict the different trajectories of A1c values across time indicated by the profiling structure.

5.2.1   Association of A1c with acute adverse events—We aim to address whether using imputed data affects inferences on how glycemic control is associated with acute diabetes outcomes. Since marginal profiling and joint profiling present similar results, we use the results under marginal profiling as an illustration. It should be noted that the MAR assumption here applies to missing A1c. The complete case analysis (CCA) applies to the occasions with collected A1c values and causes exclusion of the outcomes available at the occasions where A1c values are missing. Hence, missingness in the outcome is induced by missing A1c and may be MAR or MNAR, but imputing A1c restores all the outcomes.

The overall incidence of any acute health events is 16% across the patient-quarters. We predict this incidence by a seven-category discrete A1c indicator ($< 6\%$–reference level, 6–6.5%, 6.5–7%, 7–7.5%, 7.5–8%, 8–9%,   9%), where the levels are chosen to coincide with those used by ADA to translate A1c into blood glucose level (ADA 2018a). For simplicity, we do not add other factors. For each imputation, we estimate the incidence from logistic regression models and account for the correlation between repeated measures of the same patient by implementing generalized estimating equations with an exchangeable working correlation structure. We apply MI combining rules (Rubin 1987) to obtain the standard error estimate accounting for the uncertainty due to imputation. Figure 5.3 depicts the predicted incidence for individuals grouped by A1c levels from the four methods, and Table 5.2 presents both the point estimates and 95% CIs of the predictions. Figure 5.3 shows that the relationship between A1c values and the risk of adverse outcomes is U-shaped.

Importantly, all analyses verify that A1c at or above 8% is associated with a higher incidence of adverse events than levels of 6-7.5% as expected from ADA recommendations to maintain level below 7% or below 6.5%, if possible. In fact, imputation does not notably change means for the two categories above 8%, where less than 10% of A1c values are

imputed. Results for levels < 6%, where approximately 20% of values are imputed, are mixed, with BPMI as well as CCA showing a significantly higher incidence of acute events at this level of control than for A1c levels 6-7. 5%. While MICE and MICE-2l also show higher incidence at A1c< 6% than at the next two higher levels, these differences are not statistically significant.

The incidence between A1c of 6% and 8% shows the largest differences between CCA and analyses with imputed data. MICE and MICE-2l agree with BPMI in showing higher incidence than CCA in this range, but BPMI shows a slightly lower incidence than the other two methods, indicating that BPMI tends to place more low risk individuals in this category. The evidence from BPMI is similar to the finding in Li et al. (2019). The fact that all methods show higher incidence than CCA in this range may be due to a complex interplay of reasons for missed A1c tests and their consequences, where missed A1c may be a marker for either good and stable glycemic control, or for lower utilization of preventive care. Future work will study the heterogeneous effects of patent complexity in the nonlinear relationship.

It seems that the probability of missingness depends on the outcome in this example, so that the missingness induced by A1c is MNAR in the acute event model. However, it is less clear if explicitly taking the outcome into account would change the relationship between imputed A1c and the outcome. Doing so would require specifying a correct model for the relationship between A1c and the outcome. Including the outcome as a covariate with main effects in MICE-2l, made no difference in predictions. We posit that imputing the A1c from a rich set of repeated measures for each individual minimizes the information the outcome can bring to the imputation. It should also be noted that baseline event incidence was included as a covariate in the imputation.

5.2.2 Profiling heterogeneity and interpretation—As one important byproduct of BPMI, the latent profiling provides an interpretable summary of patient heterogeneity in A1c trajectories. As an illustration, we randomly draw the latent class indicators based on the posterior allocation probabilities and use the last draw for the profiling assignment. The three profiles have different trajectories over time and different scales. Figure 5.4 shows the estimated spline functions and 95% CIs, as well as the observed A1c values across profiles. We present the two plots in Figure 5.4 to illustrate the profiling structure, but they are not directly comparable. The spline functions ($D_i^* \beta_{C_i}^*$) are part of the mean structure, so the two

plots have different y-axis ranges. Meanwhile, we have included additional time-varying variables, and they both have significantly positive relationships with A1c values, where the coefficient estimate of age is 0.007 with 95% CI (0.006, 0.008), and the coefficient estimate of the count of physician visits is 0.003 with 95% CI (0.001, 0.006).

The profile allocation assigns 3694 patients to the largest class with the smallest variability of residuals, 0.29. The smallest class has 996 patients and the largest variability of A1c values (1.42), with large coefficients of the spline functions. The remaining class has 2692 patients, and the estimated variability of A1c values is 0.65. The patients within Profile 1 are under good control with low A1c levels and stability, of whom all the A1c values are below

9%. Patients within Profile 2 are under fair control with modest A1c values, while those with Profile 3 are under bad control with high A1c levels and variability.

The profiling is primarily dependent on the A1c trajectories, and the profile allocation probabilities are predicted by the baseline covariates. Figure 5.5 and Table 5.3 provide the descriptive summary of the baseline characteristics and the corresponding coefficients for profile allocation. The patients in poor control in Profile 3 tend to be non-white, have Medicaid coverage and more complications (such as Depression, Entitlement Disability, Eye disease, Psychoses and Obesity), and take Insulin. These patients also have more ER visits, more adverse events, higher HCC and baseline A1c values than those in good/fair control.

The coefficients for allocation across profiles give the logarithm of the relative odds to be assigned to other profiles compared to Profile 1. With higher baseline HCC and A1c values, the patients tend to be out of good control. Non-white patients with Chronic Pulmonary Disease, Entitlement Disability, other neurological disorders, Obesity, Psychoses and taking the hypoglycemic drug, have significantly higher odds of being out of good control. The patients with Dementia, Eye disease, congestive heart failure and those who take Insulin or Sulfonylureas tend to have good A1c control.

6 Discussion

Electronic health records play an increasingly important role in evidence-driven research on effective approaches to clinical therapy. However, as data are collected in an unstructured manner when patient contacts occur, EHR suffers from missing data, where for example, outcomes and predictors are not always available in the same time interval. MI is an attractive approach for filling in data needed for various analyses, allowing assessment of imputation uncertainty. However, few methods exist for filling in intermittently missing data in long time sequences. In addition, the missing mechanism is often complex raising the possibility of informative missingness. We develop a method for multiple imputation based on Bayesian latent profiling, which allows for missingness being intermittent and potentially MNAR. The method allows for a flexible mixture distribution by allowing location and scale parameters to vary across latent profiles, covariates to predict profile membership and a time trend to be modeled by splines. We compare BPMI with popular MI MAR based approaches MICE and MICE-2l by simulations in a range of settings, and find BPMI to be the most unbiased and efficient. We apply our method to missing A1c in EHRs, and find BPMI to perform well with 3 latent classes, in terms of BIC, LPML and posterior predictive checks. Hence, the theoretical properties for our approach seem to be satisfactory.

As we predict acute adverse outcomes from the EHR with and without imputation, BPMI as well as MICE and MICE-2l predicts more adverse events than does CCA in the 7-8% A1c range. Hence, CCA analysis of the EHR data cannot be trusted. While a MAR structure fits the A1c data, it appears that the missingness induced in the outcome by missing A1c is not MAR. However, all outcomes are restored when A1c is imputed. Our future work will explore including the acute outcomes in the imputation as generally recommended. However, it is unclear how much this would change the imputed values given the availability of rich A1c trajectories and many covariates. It is also challenging to specify the model to

properly account for the non-linear relationship and patient heterogeneity. Our approach avoids the need to model the dependence of the outcome on A1c, and improves the availability of the A1c values for investigating other outcomes. Importantly, our method fits a very flexible mixture model to capture the data distribution.

Several questions regarding the application of our method arise. We considered the case when only A1c values are missing and include numerous covariates. In practice, EHRs will suffer from incomplete information beyond A1c, such as the baseline variables in Table 2.1. With multivariate incomplete variables, joint imputation or sequentially conditional imputation is necessary. It is also uncertain how many latent classes are needed. BPMI makes the crucial assumption of conditional independence, given the latent classes. Likely, a greater number of classes makes this assumption more plausible, but leads to difficulties including covariates, especially those with low prevalence. Our diagnostics does not find evidence against the assumed model. Future work will investigate the specific role of different aspects of patient complexity on the relationship of A1c to acute outcomes. This will address substantive questions as well as further illuminate the plausibility of the imputations.

Several extensions of BPMI may be of interest. Besides the incorporation of time-varying health outcomes into modeling, other time-varying variables including BP, BMI, and LDL can provide additional information to impute the A1c levels. To handle their high missingness percentages, we can simultaneously model the trajectories and impute missing values via joint mixture membership profiling (Airoldi et al. 2019). The number of latent profiles can be treated as random in future work. With more information available from EHR, systematic variable selection in an integrative inference framework needs further investigation. For longitudinal data with intermittent missingness and nonlinear trajectories, it would be useful to connect our joint modeling approach with existing approaches for monotone missingness or weighting adjustments (Hogan and Daniels 2008; Robins et al. 2000).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

References

ADA (2018a, Jan). Glycemic targets. Diabetes Care 41 (1), 55–64.

ADA (2018b, Jan). Older adults. Diabetes Care 41 (1), 119–125.

Agency for Healthcare Research and Quality (2013, June). Primary care patients use interactive preventive health record integrated with electronic health record, leading to enhanced provision of preventive services.

Airoldi EM, Blei D, Erosheva EA, and Fienberg SE (2019). Handbook of Mixed Membership Models and Their Applications. CRC Press.

Allen C, LeCaire T, Palta M, Daniels K, Meredith M, and D'Alessio D (2001). Risk factors for frequent and severe hypoglycemia in type 1 diabetes. Diabetes Care 24, 1878–1881. [PubMed: 11679450]

Boyd CM, Darer J, Boult C, Fried LP, Boult L, and Wu AW (2005). Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases - implications for pay for performance. JAMA: The Journal of American Medical Association 294, 716–724.

Carpenter J and Kenward MG (2013). Multiple imputation and its application. John Wiley & Sons.

Cebul RD, Love TE, Jain AK, and Hebert CJ (2011). Electronic health records and quality of diabetes care. New England Journal of Medicine 365 (9), 825–833. [PubMed: 21879900]

Daniels MJ and Hogan JW (2008). Missing data in longitudinal studies: strategies for Bayesian modeling and sensitivity analysis. CRC Press.

Davis S and Alonso MD (2004). Hypoglycemia as a barrier to glycemic control. Journal of Diabetes Complications 18, 60–68.

Dluhy RG and McMahon GT (2008). Intensive glycemic control in the ACCORD and ADVANCE trials. New England Journal of Medicine 358, 2630–2633. [PubMed: 18539918]

Elixhauser A, Steiner C, Harris DR, and Coffey RM (1998). Comorbidity measures for use with administrative data. Medical care 36 (1), 8–27. [PubMed: 9431328]

Gelfand AE and Dey DK (1994). Bayesian model choice: asymptotic and exact calculations. Journal of the Royal Statistical Society Series B Stat. Methodol 56 (3), 501–514.

Gelman A, Hill J, Su Y-S, Yajima M, Pittau M, Goodrich B, Si Y, and Kropko J (2015). Package mi. R CRAN.

Gelman A, Jakulin A, Pittau MG, and Su Y-S (2008). A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics 2 (4), 1360–1383.

Goodman LA (2007). On the assignment of individuals to classes. Sociological Methodology 37, 1–22.

Haneuse S and Daniels M (2016). A general framework for considering selection bias in ehr-based studies: What data are observed and why? eGEMs (Generating Evidence & Methods to improve patient outcomes): 4 (1-16).

He Y, Zaslavsky AM, and Landrum MB (2010). Multiple imputation in a large-scale complex survey: A guide. Statistical Methods in Medical Research 19, 653–670. [PubMed: 19654173]

Hogan JW and Daniels MJ (2008). Missing Data in Longitudinal Studies. Boca Raton: Chapman and Hall.

Ibrahim JG, Chen MH, Lipsitz SR, and Herring AH (2005). Missing data methods for generalized linear models: A comparative review. Journal of the American Statistical Association 100, 332–346.

Ibrahim JG and Molenberghs G (2009). Missing data methods in longitudinal studies: a review. Test 18 (1), 1–43. [PubMed: 21218187]

Jackson GL, Edelman D, and Weinberger M (2006). Simultaneous control of intermediate diabetes outcomes among veterans affairs primary care patients. Journal of the General Internal Medicine 21, 1050–1056.

Kennedy RL, Henry J, Chapman A, Nayar R, Grant P, and Morris AD (2002). Accidents in patients with insulin-treated diabetes: Increased risk of low-impact falls but not motor vehicle crashes-A prospective register-based study. The Journal of Trauma-Injury Infection & Critical Care 52, 660–666.

King G, Honaker J, Joseph A, and Scheve K (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. American Political Science Review 95, 49–69.

Li F-R, Zhang X-R, Zhong W-F, Li Z-H, Gao X, Kraus VB, Lv Y-B, Zou M-C, Chen G-C, Chen P-L, Zhang M-Y, Kur AKA, Shi X-M, Wu X-B, and Mao C (2019). Glycated hemoglobin and

all-cause and cause-specific mortality among adults with and without diabetes. The Journal of Clinical Endocrinology & Metabolism online.

Lin H, McCulloch CE, and Rosenheck RA (2004). Latent pattern mixture models for informative intermittent missing data in longitudinal studies. Biometrics 60, 295–305. [PubMed: 15180654]

Little RJA (1995). Modeling the drop-out mechanism in repeated-measures studies. Journal of the American Statistical Association 90 (431), 1112–1121.

Little RJA and Rubin DB (2002). Statistical Analysis with Missing Data: Second Edition. New York: John Wiley & Sons.

Martin CL, Albers J, Herman WH, Cleary P, Waberski B, Greene DA, Stevens MJ, and Feldman EL (2006). Neuropathy among the diabetes control and complications trial cohort 8 years after trial completion. Diabetes Care 29, 340–344. [PubMed: 16443884]

McCulloch CE, Neuhaus JM, and Olin RL (2016). Biased and unbiased estimation in longitudinal studies with informative visit processes. Biometrics 72, 1315–1324. [PubMed: 26990830]

McFarlane SI, Jacober SJ, N NW, J JK, Castro JP, Wui MA, Gliwa A, Von Gizycki H, and Sowers JR (2002). Control of cardiovascular risk factors in patients with diabetes and hypertension at urban academic medical centers. Diabetes Care 25, 718–723. [PubMed: 11919131]

Meng XL (1994). Posterior predictive p-values. Annals of Statistics 22, 1142–1160.

National Research Council (2010). The prevention and treatment of missing data in clinical trials. Washington, D.C: The National Academies Press.

Niefeld MR, Braunstein JB, Wu AW, Saudek CD, Weller WE, and Anderson GF (2003). Preventable hospitalization among elderly medicare beneficiaries with type 2 diabetes. Diabetes Care 26, 1344–1349. [PubMed: 12716786]

Pogach LM, Tiwari A, Maney M, Rajan M, Miller DR, and Aron D (2007). Should mitigating comorbidities be considered in assessing healthcare plan performance in achieving optimal glycemic control? The American Journal of Managed Care 13, 133–140. [PubMed: 17335354]

Polson NG, Scott JG, and Windle J (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. Journal of the American Statistical Association 108 (504), 1339–1349.

Qaseern A, Vijan S, Snow V, Cross JT, Weiss KB, and Owens DK (2007). Glycemic control and type 2 diabetes mellitus: The optimal hemoglobin a1c targets - A guidance statement from the american college of physicians. Annals of Internal Medicine 147, 417–422. [PubMed: 17876024]

Raghunathan TE, Lepkowski JM, van Hoewyk J, and Solenberger P (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology 27, 85–96.

Robins JM, Hernán MA, and Brumback B (2000). Marginal structural models and causal inference in epidemiology. Epidemiology 11 (5), 550–560. [PubMed: 10955408]

Rodríguez A and Dunson DB (2011). Nonparametric bayesian models through probitstickbreaking processes. Bayesian Analysis 6 (1), 145–178.

Rubin DB (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons.

SAS Institute Inc. (2017). SAS/STAT 9.4 User's Guide.

Schafer JL (1997). Analysis of Incomplete Multivariate Data. London: Chapman & Hall.

Schafer JL (2016). Multiple imputation for multivariate panel or clustered data. https://cran.r-project.org/web/packages/pan/.

Schech S, Graham D, Staffa J, Andrade SE, La Grenade L L, Burgess M, Blough D, Stergachis A, Chan KA, Platt R, and Shatin D (2007). Risk factors for statin-associated rhabdomyolysis. Pharmacoepidemiol Drug Safety 16, 352–358.

Shorr RI, Ray WA, Daugherty JR, and Griffin MR (1997). Incidence and risk factors for serious hypoglycemia in older persons using insulin or sulfonylureas. Archives of Internal Medicine 157, 1681–1686. [PubMed: 9250229]

Si Y and Reiter JP (2013). Nonparametric bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. Journal of Educational and Behavioral Statistics 38 (5), 499–521.

Si Y, Reiter JP, and Hillygus DS (2016). Bayesian latent pattern mixture models for handling attrition in panel studies with refreshment samples. The Annals of Applied Statistics 10, 118–143.

Sun B and Tchetgen EJT (2016). On inverse probability weighting for nonmonotone missing at random data. Journal of the American Statistical Association 113 (521), 369–379.

The Centers for Disease Control and Prevention (2017). National diabetes statistics report. https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf.

U.S. Center for Medicare & Medicaid Services (2012, Mar). Electronic health records. https://www.cms.gov/Medicare/E-Health/EHealthRecords/index.html.

Van Buuren S and Oudshoorn C (1999). Flexible multivariate imputation by MICE. Technical report, Leiden: TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054.

Vermunt JK (2010). Latent class modeling with covariates: Two improved three-step approaches. Political Analysis 18, 450–468.

Wu MC and Carroll RJ (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. Biometrics 44 (1), 175–188.

Zeldow B, Flory J, Stephens-Shields A, Raebel M, and Roy J (2018). Outcome identification in electronic health records using predictions from an enriched Dirichlet process mixture. https://arxiv.org/abs/1806.02411.
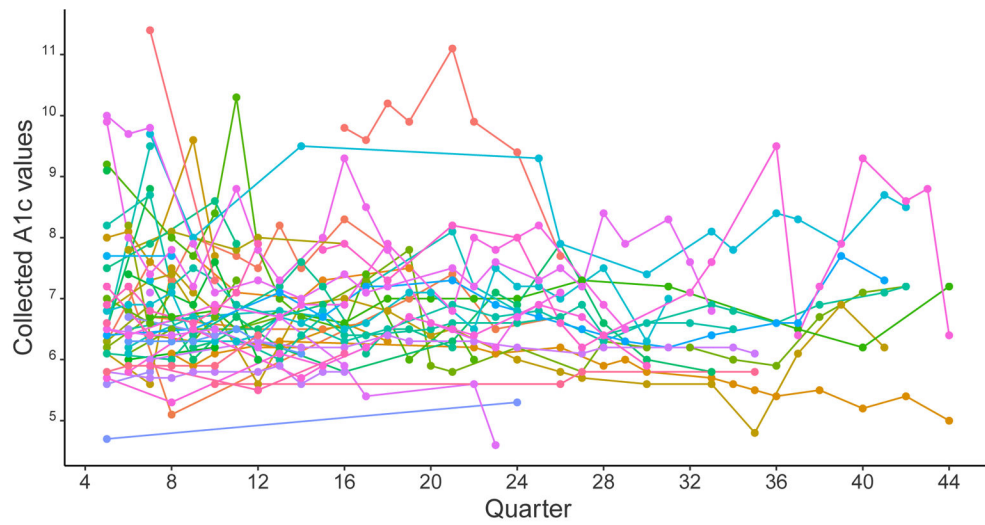
Figure 2.1:
Observed A1c values of 30 randomly selected patients. Each line represents one patient, with segments connecting two available subsequent A1c measurements.
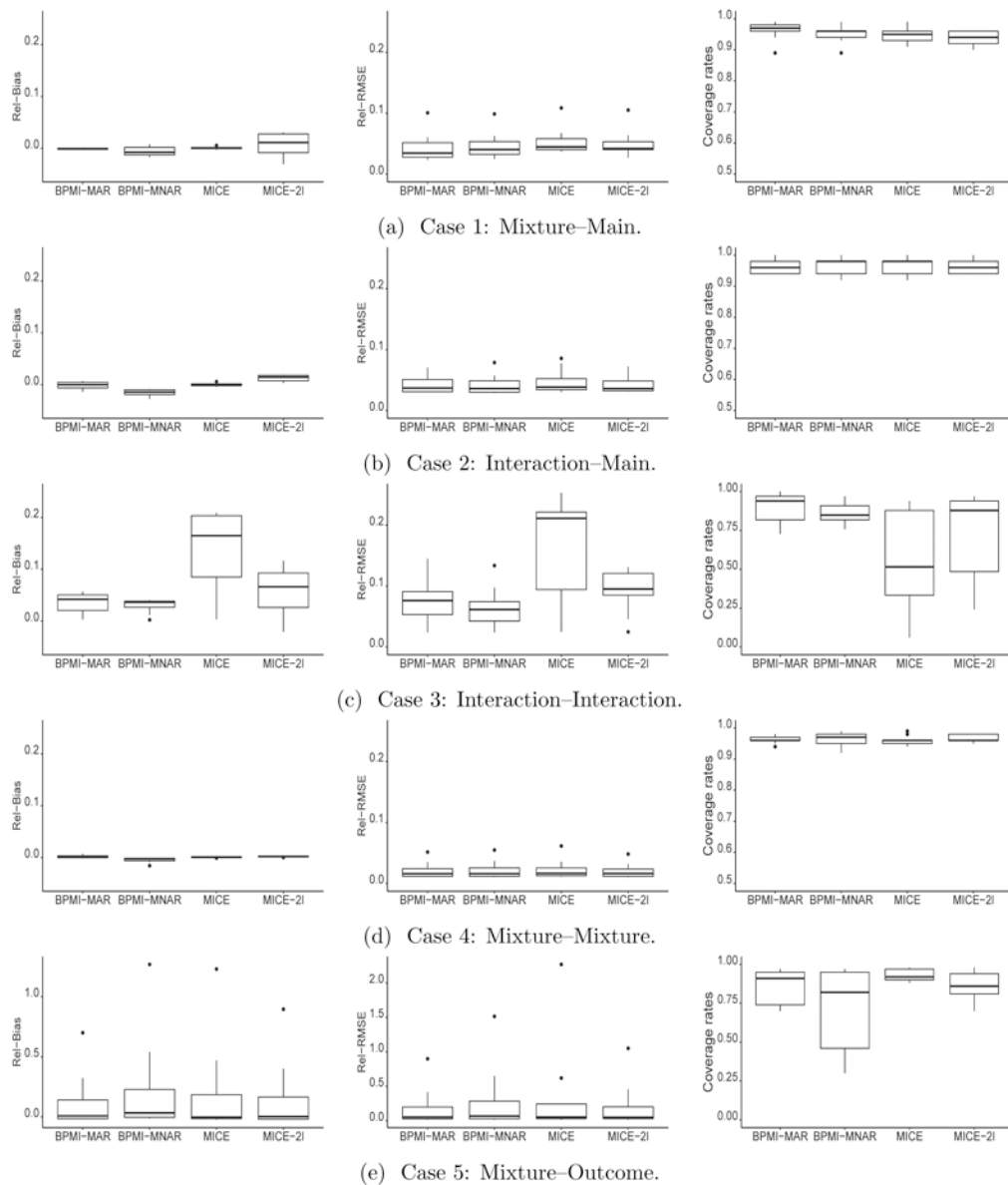
(a) Case 1: Mixture–Main.

(b) Case 2: Interaction–Main.

(c) Case 3: Interaction–Interaction.

(d) Case 4: Mixture–Mixture.

(e) Case 5: Mixture–Outcome.

Figure 4.1:
Performance of MI inferences on the average outcome values at 9 follow-up occasions with 10 imputed datasets comparing different approaches: marginal profiling (BPMI-MAR), joint profiling (BPMI-MNAR), sequential imputation MICE and multilevel imputation MICE-2l.
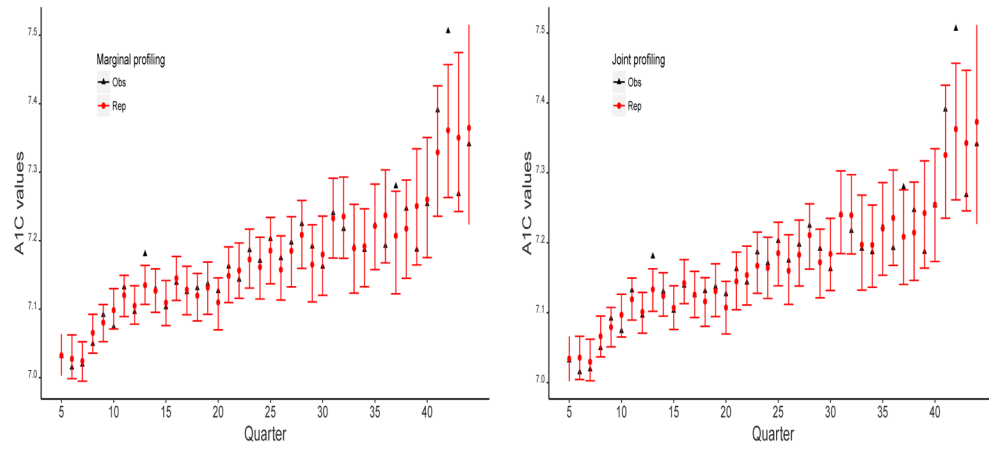
Figure 5.1:
Posterior predictive check on quarter-level summaries for marginal and joint profiling approaches. The black triangles represent the means of observed data, the red dots are the posterior predictive mean estimates, and the red error bars are the 95% credible intervals of the predictive means.
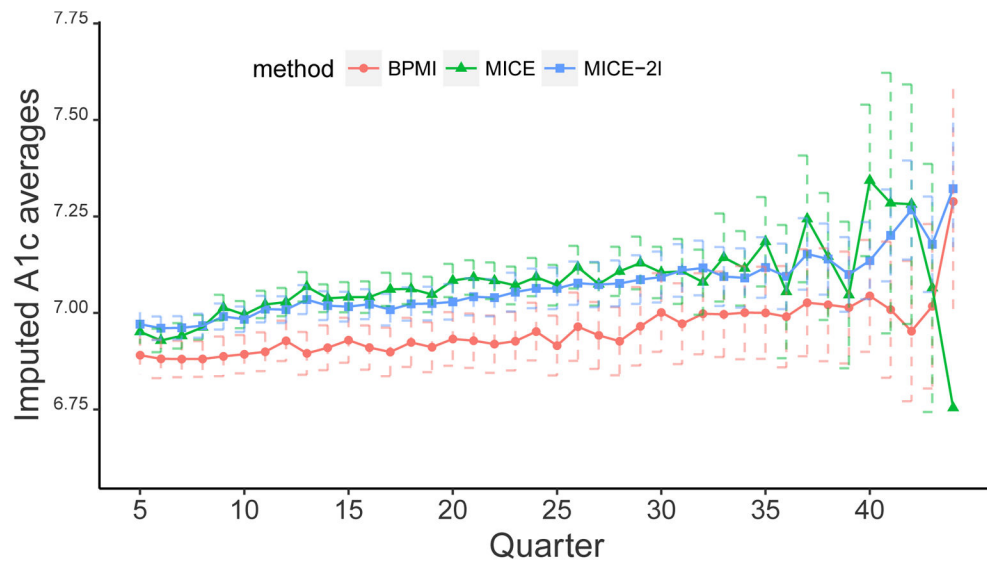
Figure 5.2:
Comparison of the average of imputed A1c values and 95% confidence intervals obtained by BPMI, MICE and MICE-2l across follow-up quarters. The wide CI of the MICE output for the last quarter (6.11, 7.40) is omitted.
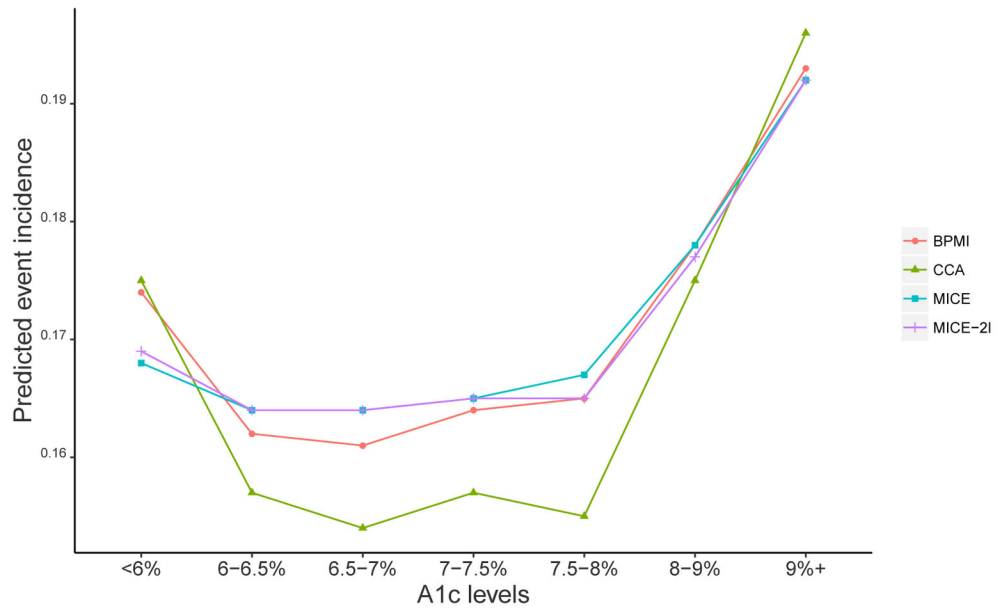
Figure 5.3:
Comparison between MI inferences and complete case analysis (CCA) in predictive acute health event incidence risk.
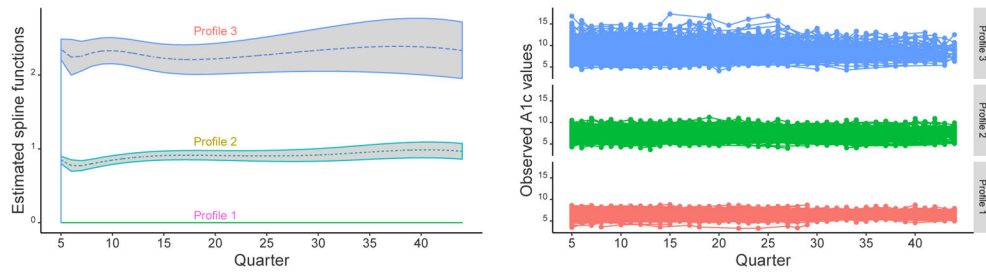
Figure 5.4:
Estimated spline functions with 95% confidence intervals and observed A1c values of patients across profiles.

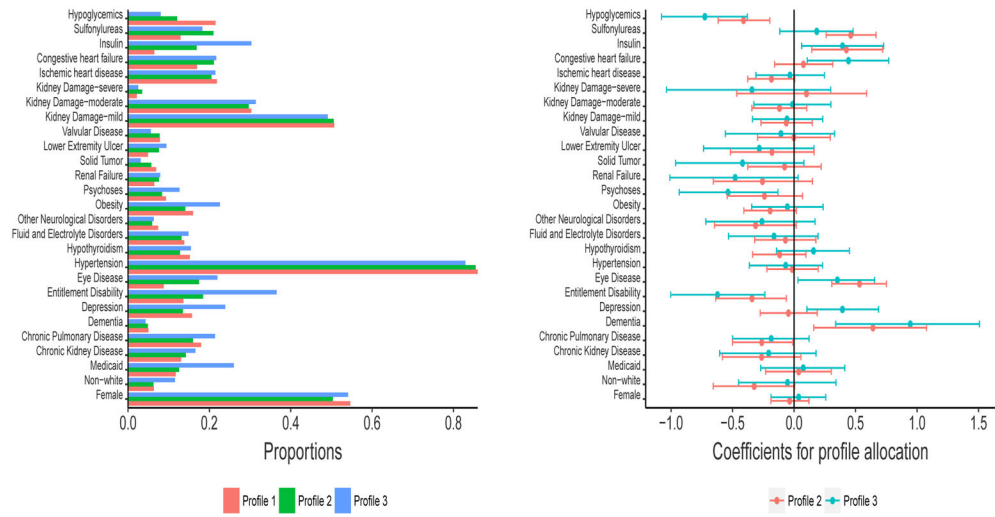**Figure 5.5:**
Frequency distribution of categorical baseline characteristics and their coefficients for profile allocation.

Table 2.1:

Summary of baseline characteristics: mean (standard deviation) and percentage.

Sociodemographics	
Age	69.91 (10.76)
Female	53%
White	93%
Medicaid	14%
Utilization	
Adverse event count	0.58 (0.88)
Hierarchical condition categories	1.41 (0.98)
Number of emergency room visits	0.12 (0.37)
Comorbidity and complication	
Count of low prevalence conditions	0.21 (0.54)
Chronic kidney disease	14%
Chronic pulmonary disease	18%
Dementia	5%
Depression	16%
Entitlement disability	18%
Eye disease	14%
Hypertension	85%
Hypothyroidism	14%
Fluid and electrolyte disorders	14%
Other neurological disorders	7%
Obesity	16%
Psychoses	9%
Renal failure	7%
Solid tumor without metastasis	6%
Lower extremity ulcer	7%
Valvular disease	8%
Kidney damage-light	16%
Kidney damage-mild	51%
Kidney damage-moderate	30%
Kidney damage-severe	3%
Congestive heart failure	19%
Ischemic heart disease	21%
Medication prescribed	
Insulin	13%
Sulfonylureas	17%
Hypoglycemics	16%
Other	54%

Table 5.1:

Comparison of models with different number of profiles (*L*).

Marginal profiling			Joint profiling		
L	BIC	LPML	L	BIC	LPML
2	126857	−106129	2	**323529**	−192506
3	**124145**	**−94305**	3	329037	**−174286**
4	130872	−97579	4	334345	−180697
5	136263	−97315			
6	137082	−97397			

Table 5.2:

Predictive acute health event incidence risk and 95% confidence intervals for individuals grouped by A1c levels.

	BPMI	**MICE**	**MICE-2l**	**CCA**
< 6%	0.174 (0.164, 0.182)	0.168 (0.16, 0.177)	0.169 (0.161, 0.178)	0.175 (0.164, 0.187)
6 – 6.5%	0.162 (0.156, 0.169)	0.164 (0.157, 0.169)	0.164 (0.157, 0.171)	0.157 (0.149, 0.164)
6.5 – 7%	0.161 (0.154, 0.168)	0.164 (0.157, 0.169)	0.164 (0.157, 0.169)	0.154 (0.147, 0.161)
7 – 7.5%	0.164 (0.157, 0.172)	0.165 (0.158, 0.172)	0.165 (0.158, 0.172)	0.157 (0.149, 0.165)
7.5 – 8%	0.165 (0.157, 0.175)	0.167 (0.157, 0.175)	0.165 (0.156, 0.174)	0.155 (0.146, 0.165)
8 – 9%	0.178 (0.169, 0.187)	0.178 (0.169, 0.187)	0.177 (0.167, 0.185)	0.175 (0.164, 0.185)
9%+	0.193 (0.179, 0.206)	0.192 (0.179, 0.206)	0.192 (0.179, 0.206)	0.196 (0.182, 0.211)

Table 5.3:

Descriptive statistics of baseline characteristics (first 4 quarters) and their coefficients for allocation across profiles.

	Profile 1	Profile 2	Profile 3
Descriptive statistics: mean (sd)			
#low prevalence	0.22 (0.54)	0.21 (0.54)	0.21 (0.54)
#ER visits	0.11 (0.39)	0.10 (0.29)	0.17 (0.45)
#Adverse events	0.54 (0.85)	0.57 (0.86)	0.73 (0.97)
HCC	1.26 (0.84)	1.57 (1.11)	1.57 (1.01)
Age	71.52 (9.98)	69.76 (10.28)	64.33 (12.73)
A1c	6.38 (0.63)	7.23 (0.93)	8.5 (1.66)
Coefficients: log-odds (95% CI)			
Intercept	0	−6.08 (−7.51, −4.94)	−9.15 (−10.29, −8.01)
#low prevalence	0	−0.34 (−0.52, −0.16)	−0.44 (−0.68, −0.21)
#ER visits	0	−0.46 (−0.81, −0.13)	−0.41 (−0.79, −0.05)
#Adverse events	0	0.12 (−0.01, 0.25)	0.15 (−0.02, 0.32)
HCC	0	0.58 (0.36, 0.73)	0.59 (0.36, 0.76)
Age	0	−0.06 (−0.07, −0.04)	−0.11 (−0.12, −0.09)
Baseline A1c	0	1.35 (1.25, 1.46)	2.06 (1.92, 2.19)